DOCUMENT RESUME

ED 292 888 TM 011 355

AUTHOR Kingston, Neal M.; McKinley, Robert L.
TITLE Assessing the Structure of the GRE General Test Using Confirmatory Multidimensional Item Response Theory.
PUB DATE 1 Apr 88
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Factor Analysis; Higher Education; Item Analysis; *Latent Trait Theory; Scoring; *Statistical Analysis; *Test Format; Test Interpretation
IDENTIFIERS *Graduate Record Examinations

ABSTRACT

Confirmatory multidimensional item response theory (CMIRT) was used to assess the structure of the Graduate Record Examination General Test, about which much information about factorial structure exists, using a sample of 1,001 psychology majors taking the test in 1984 or 1985. Results supported previous findings that, for this population, there exists a weak analytical factor defined by the logical reasoning items and not by the analytical reasoning items. This finding was more straightforward than the same finding based on the full-information factor analysis (FIFA) approach used in the previous study; however, an advantage of FIFA was that it allowed the researcher to assess the proportion of variance explained by each factor in the orthogonal solution--the CMIRT approach is unlikely to yield such a statistic. Examination of the results shows the CMIRT approach to hold much promise. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED292888

Assessing the Structure of the GRE General Test Using Confirmatory Multidimensional Item Response Theory[1]

Neal M. Kingston
Robert L. McKinley
Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

NEAL M. KINGSTON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

[1] Paper presented as part of the symposium, "Item response theory meets multidimensional tests," at the annual meeting of the American Educational Research Association, New Orleans, April 8, 1988.

TM 011 355

INTRODUCTION

Assessing the Dimensionality of Binary Data

It is often important to assess the underlying factorial structure of mental test data. For example, such analyses are a useful step in the construct validation of a test or battery. Typically a factor analytic approach is used. First the inter-item correlation matrix is computed. Depending on the purpose of the analysis and the existing body of knowledge regarding the test and the constructs it measures, the researcher will select one of two major factor analytic approaches--exploratory or confirmatory. Under either approach, one or more factor analytical models are fit to the data, compared, and the researcher decides which model best fits the data.

Whether the researcher uses an exploratory or confirmatory approach, the factor analysis of mental test data is fraught with difficulties. These difficulties stem primarily from statistical artifacts associated with estimating the intercorrelations among binary data. These difficulties include:

> mismatch of assumptions of phi coefficients with underlying data (i.e, underlying data are continuous, not binary);
>
> mismatch of assumptions of tetrachoric correlations with underlying data (i.e., underlying data are not based on normal distributions and item responses are affected by guessing; and
>
> the appearance of difficulty factors whether the researcher uses phi or tetrachoric correlations.

For a more full explanation of the difficulties associated with assessing the dimensionality of binary data, the interested reader is referred to Lord (1980, p20) and Mislevy (1986).

To overcome these difficulties, Bock, Gibbons, and Muraki (1985) developed a new approach to exploratory factor analysis called full-information factor analysis (FIFA). FIFA, as implemented in the computer program TESTFACT (Wilson, Wood, and Gibbons, 1984), uses the marginal maximum likelihood method (Bock and Aitkin, 1981) to estimate reparameterized discrimination and difficulty parameters for multidimensional item response theory (IRT) models. The IRT parameter estimates are then used to estimate the interitem correlation matrix, which is used as the basis for a principal factors analysis.

One particularly attractive feature of TESTFACT is that it can be used to perform a stepwise analysis. To do this, you sequentially estimate parameters for higher and higher order multidimensional IRT

models. A likelihood ratio chi-square test is used to compare the likelihood of the parameter set, given the observed data, for the one dimensional IRT model with that of the two dimensional model. Then the likelihoods of the two and three dimensional models are compared. Likelihoods of higher and higher order IRT models can be compared for as long as the researcher can afford (parameter estimation in multidimensional IRT models is time consuming and/or expensive; time is related exponentially to the number of dimensions). Note, it is the likelihoods based on the IRT model that are used to perform a test of statistical significance on the additional factors, not data from the factor analytic model.

The Structure of the GRE General Test

The GRE General Test consists of seven sections administered in two-and-one-half hours. Examinee responses to items on six of the sections count toward an examinees scores: two sections each toward the verbal, quantitative, and analytical scores. Each section within a pair is developed to be statistically and content parallel to the other section. The seventh section (which might be in any position within the test) does not count toward the examinee's score. Instead, it is used typically to pretest items to ensure that future editions of the test are of high quality. In other cases the seventh section is used to try out new item types or for other experimental purposes.

Each measure consists of two or more item types that are intended to tap the underlying construct of interest. Table 1 presents the number of items of each item type in one section of each measure and the number of choices per item.

Table 1
Item types in the GRE General Test

Item Type	Number of Items	Choices Per Item
Verbal	38	
Analogies	9	5
Antonyms	11	5
Reading Comprehension	11	5
Sentence Completion	7	5
Quantitative	30	
Data Interpretation	5	5
Discrete Quantitative	10	5
Quantitative Comparisons	15	4
Analytical	25	
Analytical Reasoning	19	5
Logical Reasoning	6	5

Stricker and Rock (1985) performed confirmatory factor analyses on the GRE General Test using a correlation matrix based on item parcels. Of the solutions they tried, the one that provided the best fit to the data was the one that matched the GRE score reporting scheme. That is, the verbal items all loading on one factor, the quantitative items all on a second factor, and the analytical items on a third factor. However, they found that the analytical items all had noticeably lower loadings on the hypothesized analytical factor than did the verbal or quantitative items on their respective factors. In addition, the analytical factor correlated considerably more highly with the verbal and quantitative factors than the verbal and quantitative factors did with each other (an average of .79 across three different samples compared to an average of .58).

Kingston (1984) found that when scores based on item type were intercorrelated, the analytical reasoning item type correlated more highly with each of the four verbal item types than it did with logical reasoning. Also, the reading comprehension and analytical items correlated more highly with each other than either did with any of the other three verbal item types. On the other hand, logical reasoning scores correlated more highly with each of the quantitative item types than it did with analytical reasoning. This suggests that a model that proposed only a verbal and a quantitative factor, where logical reasoning items were allowed to load on the quantitative factor and analytical reasoning items were allowed to load on the verbal factor, might fit GRE General Test data better than any of the models tried by Stricker and Rock.

In a study of the incremental validity of the analytical measure, Kingston (1985) found evidence suggesting that the factor structure of the GRE might be different for subpopulations with different undergraduate majors. Based on this hypothesis of differential factor structure and the availability of a new and theoretically superior exploratory factor analytic approach--full-information factor analysis, Schaeffer and Kingston (1988) analyzed the GRE factor structure for seven samples of GRE examinees (three randomly equivalent groups of psychology majors, and one group each of education, engineering, English, and mathematics majors). They found evidence of a relatively weak but statistically significant analytical factor for all groups analyzed with the possible exception of education majors. However, differences in strength and order of factors extracted in the three randomly equivalent groups of psychology majors was disturbing and cast doubt on the comparisons of the relative strength of the factors in the other groups. Such differences might easily occur in an exploratory solution if the likelihood surface were relatively flat. Due to sampling error exacerbated by the correlations among factors as well as the large number of parameters that need to be estimated in a multidimensional IRT solution, many different sets of parameter estimates might fit the data almost equally well. The chosen solution might have little relation to a structure supported by psychological theory, although alternative theoretically parsimonious structures that fit essentially as well might exist.

Purpose of This Study

A new model, confirmatory multidimensional item response theory (CMIRT), has been developed to address the problems that have been observed with other approaches to assessing the dimensionality of mental test data. The purpose of this paper is to explore the use of this model by applying it to a test for which much information about factorial structure exists--the GRE General Test.

Confirmatory Multidimensional Item Response Theory

To avoid the potential problems inherent in an exploratory approach such as FIFA, McKinley and Kingston (1988) developed a new IRT-based dimensionality assessment method--confirmatory multidimensional item response theory (CMIRT). The referenced paper describes in detail the model and estimation procedures, and thus we will not spend much time describing the model here.

The CMIRT model is a variant of the same multidimensional IRT model used in TESTFACT (Wilson, Wood, and Gibbons, 1984), MULTIDIM (McKinley, 1987) and elsewhere. This basic model is:

$$P_i(\underline{\theta}_j) = c_i + (1-c_i)/(1+\exp(-D(b_i + \underline{a}_i'\underline{\theta}_j))), \qquad (1)$$

where $P_i(\underline{\theta}_j)$ is the probability of a correct response by an examinee j,

$\underline{\theta}_j$ is the ability parameter vector for examinee j,

c_i is the lower asymptote parameter for item i,

D is a scaling constant approximately equal to 1.702,

b_i is the threshold parameter for item i, and

$\underline{a}_i$ is a vector of discrimination parameters for item i.

The ability and discrimination parameter vectors contain one element for each dimension in the hypothesized model.

The CMIRT model differs from the basic multidimensional IRT model in that a structure matrix is used to impose constraints on the item discrimination parameters. As implemented in the program CONFIRM, the CMIRT model requires all items to load on a first or general factor. Every item can load also on any <u>one</u> of the hypothesized second order factors. The structure matrix is used to indicate for which dimensions a discrimination parameter will be estimated.

For example, if you had a four item science test with two physics items followed by two chemistry items, either of the following two structures might apply:

$$\underline{S}(1) = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} \quad , \tag{2}$$

$$\underline{S}(2) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad . \tag{3}$$

Structure one hypothesizes a single general science factor. Structure two hypothesizes a general science factor and two specific factors--physics and chemistry.

DATA

Data from Psychology Majors sample number two from Schaeffer and Kingston (1988) were used for all analyses. Examinees were selected from the approximately 82,000 examinees who took a particular edition of the test in October 1984, April 1985, or December 1985 and who indicated that English was their best language, that they had not previously taken any GRE test, that they were in their senior year in college when they took the test, and that there undergraduate major was psychology. Of the 3,325 examinees who fit this definition, 1,001 were selected randomly for this sample.

CMIRT analyses are computationally intensive. In order to perform analyses within a reasonable time frame, and in order to perform analyses on the same data used by Schaeffer and Kingston so as to be able to compare results, only 93 of the 186 items (one half of each item type) on the GRE General Test were included in the analyses. This was done by selecting the first of each pair of separately timed parallel sections of verbal, quantitative, and analytical items. The number of items of each type was presented previously in Table 1. Items were scored 1 if correct and 0 if wrong or omitted, in keeping with the number-right scoring instructions of the General Test.

Table 2 presents the means and standard deviations of the raw scores on each of the three General Test measures. The data indicate that the test was of approximately middle difficulty for this sample.

Table 2
Performance of Sample
on Verbal, Quantitative, and Analytical Sections

	V	Q	A
n items	38	30	25
mean	21.7	17.0	13.6
s.d.	5.4	4.6	3.6

MODELS

Eight different models were fit to the data. First, three different unidimensional IRT models were fit to the data: the one- (1D-1PL), two- (1D-2PL), and three-parameter (1D-3PL) logistic models. This was done to select a model whose likelihood would then be used as a baseline for comparing the multidimensional models, as previous studies indicated one dimension would not be able to adequately explain the data. In the one- and two-parameter models all lower asymptotes were set to the lower of .15 or .9 times the proportion correct rather than the value zero used by some other researchers. This choice reflects the non-zero probability of a correct response that has long been observed with multiple-choice items.

Estimating accurately the lower asymptote is difficult even when fitting a one-dimensional model (see, e.g., Thissen and Wainer, 1982). It is likely to be more difficult yet to estimate c with a multidimensional model. To avoid these problems at this early stage of experimentation with CMIRT, we have chosen to use a modified two-parameter model for all multidimensional models in this study. Therefore, for each of the five models with higher dimensionality, a constant non-zero lower asymptote, as described above, was used.

The five multidimensional models fit to the GRE data follow:

3D-G,V,Q -- 3 dimensions: general, verbal (all verbal items plus logical reasoning items), and quantitative (all quantitative items plus analytical reasoning items);

4D-G,V,Q,A -- 4 dimensions: general, verbal, quantitative, and analytical;

4D-G,DV,RC,Q -- 4 dimensions: general, discrete verbal (analogies, antonyms, and sentence completion), reading comprehension (reading comprehension and logical reasoning), and quantitative (all quantitative items plus analytical reasoning);

4D-G,V,Q,LR -- 4 dimensions: general, verbal, quantitative (all quantitative items plus analytical reasoning), and logical reasoning); and

4D-G,R_1,R_2,R_3 -- 4 dimensions: general, and three specific factors formed by alternately assigning items to the first, second, and third specific factor regardless of content or item type (i.e., three random factors).

RESULTS

Table 3 presents the results for the three unidimensional models, -2 log likelihood, degrees of freedom, the chi square of the difference between the -2 log likelihoods of that model and the one dimensional three-parameter logistic model, the degrees of freedom for that difference, and the probability of that difference occurring by chance.

Table 3
Comparison of Unidimensional Models

Model[1]	-2 log likelihood	NPE[2]	chi square difference[3]	df	p
1D-1PL	98,750.2	93			
1D-2PL	97,074.4	186	1,675.8	93	<.0001
1D-3PL	96,982.7	224[4]	91.7	38	<.0001

[1] See text for description of models.

[2] Number of parameters estimated.

[3] The difference between -2 log likelihoods for subsuming models is distributed chi square with degrees of freedom equal to the difference in the number of parameters estimated. The chi-square presented in this table tests the null hypothesis that a model fits the data no better than the model on the preceding line of the table.

[4] 56 items had their c set to a common value.

In comparing the results of the one factor models, it must be remembered that a one dimensional model was applied to data that were definitely multidimensional. The results show that in this sample of psychology majors, the two-parameter logistic model fit the data much better than the one-parameter model. That is, the likelihood of the data given the two-parameter model is approximately 10^{838} times greater than given the one-parameter model. The three-parameter model fit somewhat better than the two-parameter model; the difference in fit was statistically significant at beyond the .0001 level.

Table 4 repeats the results for the 1D-3PL model and presents the results for the five multidimensional models. The 1D-3PL model is repeated to make it easy to compare to its likelihood the likelihoods of the various multidimensional models. The structure of the table is the same as Table 3 except that the chi-square difference tests for each multidimensional model compares that model with the 1D-3PL model.

Table 4
Comparison of Multidimensional Models with 1D-3PL Model

Model[1]	-2 log likelihood	NPE[2]	chi square difference[3]	df	p
1D-3PL[4]	96,982.7	224[4]			
3D-G,V,Q	96,386.8	279	595.9	55	<.0001
4D-G,V,Q,A	96,268.7	279	714.0	55	<.0001
4D-G,V,Q,LR	96,246.0	279	736.7	55	<.0001
4D-G,DV,RC,Q	96,362.9	279	619.8	55	<.0001
4D-G,R_1,R_2,R_3	97,459.8	279	-477.1	55	

[1] See text for description of models.

[2] Number of parameters estimated.

[3] The difference between -2 log likelihoods for subsuming models is distributed chi square with degrees of freedom equal to the difference in the number of parameters estimated. The chi-square presented in this table tests the null hypothesis that a model fits the data no better than the model on the preceding line of the table.

[4] 56 items had their c set to a common value.

The three-factor solution with logical reasoning items included with the verbal item types and analytical reasoning items included with the quantitative item types yielded a large improvement in fit over the one-factor three-parameter model. This is not surprising as the existence of verbal and quantitative factors within the GRE General Test is well documented (Powers and Swinton, 1981; Powers, Swinton, and Carlson, 1977; Rock, Werts, and Grandy, 1982; Schaeffer and Kingston, 1988; Stricker and Rock, 1985; Swinton and Powers, 1980).

The 4D-G,V,Q,A model matches the nominal structure of the General Test (i.e., verbal items go together on a verbal factor, quantitative items on a quantitative factor, and analytical items on an analytical factor). The likelihood of this model given the observed data is 10^{59} times greater than that of the 3D-G,V,Q model. This result suggests that the verbal and quantitative factors alone are not sufficient to describe the items in the current GRE General Test.

The second of the four dimensional models, 4D-G,DV,RC,Q was suggested by the correlational evidence presented by Kingston (1984). However, at least in this sample of psychology majors this model fit less well than the G,V,Q,A model.

The third four dimensional model, 4D-G,V,Q,LR was suggested by the results of Schaeffer and Kingston (1988). In this model the analytical factor is defined by only the logical reasoning item type; analytical reasoning items are put into the quantitative factor. This model had the greatest likelihood of any of the structures fit to the data in this research.

The final model applied to these data, 4D-R_1,R_2,R_3, had a lower likelihood than any model other than the 1D-1PL model. The failure of this model makes sense as there is no reason to expect the items within each of the three random factors to hang together better than they do with items in the alternative random factors.

CONCLUSIONS

The CMIRT analyses of the GRE General Test based on the one sample of psychology majors supports the previous findings of Schaeffer and Kingston: for this population there exists a weak analytical factor defined by the logical reasoning items and not the analytical reasoning items. Using the confirmatory approach in this study, this finding was much more straight-forward than the same finding based on the exploratory full-information factor (FIFA) analysis approach used by Schaeffer and Kingston. One advantage of the FIFA approach, however, was it allowed the researcher to assess the proportion of variance explained by each factor in the orthogonal solution. A similar statistic based on the CMIRT approach appears unlikely.

Clearly, more research on CMIRT is necessary. The results using real data presented here, combined with the simulation results presented by McKinley and Kingston (1988), indicate the CMIRT approach holds much promise.

A NOTE ON ESTIMATION

All analyses using the program CONFIRM were run on either an IBM XT running at 4.7 megahertz with an 8087 numerical coprocessor or a Compaq 386 portable running at 20 megahertz using an 80387 numerical coprocessor. Using the Compaq 386/20, the 1D-3PL model required 4.5 minutes per estimation cycle and took about 15 cycles to achieve acceptable convergence. The 4D-G,V,Q,A model required about 38 minutes per cycle and achieved acceptable convergence in 14 cycles. Analyses on the IBM XT required approximately 9.5 times longer to run.

REFERENCES

Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. Psychometrika, 46, 443-445.

Bock, R.D., Gibbons, R., and Muraki, E. (1985). Full-information factor analysis. MRC Report 85-1. Chicago: National Opinion Research Center.

Kingston, N.M. (1984). Reanalysis of the psychometric characteristics of the revised analytical measure of the GRE General Test. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Kingston, N.M. (1985). The incremental validity of the GRE analytical measure for various department types. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

McKinley, R.M. (1987). Users guide to MULTIDIM. Princeton, NJ: Educational Testing Service.

McKinley, R.M. and Kingston, N.M. (1988, April). Confirmatory analysis of test structure using multidimensional item response theory. Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, LA.

Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Powers, D.E. and Swinton, S.S. (1981). Extending the measurement of graduate admission abilities beyond the verbal and quantitative domains. Applied Psychological Measurement, 5, 141-158.

Powers, D.E., Swinton, S.S., and Carlson, A.B. (1977). A factor analysis of the GRE Aptitude Test. GRE Board Professional Report No. 75-11P. Princeton, NJ: Educational Testing Service.

Rock, D., Werts, and Grandy, J. (1982). Construct validity of the GRE Aptitude Test across populations--An empirical confirmatory study. GRE Board Professional Report No. 78-1P. Princeton, NJ: Educational Testing Service.

Schaeffer, G.A. and Kingston, N.M. (1988). Strength of the analytical factor in several subgroups: A full-information factor analysis approach. GRE Board Professional Report No. 86-7P. Princeton, NJ: Educational Testing Service.

Stricker, L. J. and Rock, D.A. (1985). Factor structure of the GRE General Test for older examinees: Implications for construct validity.

GRE Board Research Report No. 83-10R. Princeton, NJ: Educational Testing Service.

Swinton, S.S. and Powers, D.E. (1980). A factor analytic study of the restructured GRE Aptitude Test. GRE Board Professional Report No. 77-6P. Princeton, NJ: Educational Testing Service.

Thissen D. and Wainer, H. (1982). Some standard errors in item response theory. Psychometrika, 47, 397-412.

Wilson, D., Wood, R.L., and Gibbons, R. (1984). TESTFACT: Test scoring and item factor analysis [computer program]. Chicago, IL: Scientific Software.